Human Molecular Genetics, 2009, Vol. 18, Review Issue 2 **R163–R168**
doi:10.1093/hmg/ddp396

Cancer genome sequencing: a review

Elaine R. Mardis* **and Richard K. Wilson**

The Genome Center, Department of Genetics, Washington University School of Medicine, St. Louis, MO 63108, USA

Received July 5, 2009; Revised and Accepted August 17, 2009

A genomic era of cancer studies is developing rapidly, fueled by the emergence of next-generation sequencing technologies that provide exquisite sensitivity and resolution. This article discusses several areas within cancer genomics that are being transformed by the application of new technology, and in the process are dramatically expanding our understanding of this disease. Although, we anticipate that there will be many exciting discoveries in the near future, the ultimate success of these endeavors rests on our ability to translate what is learned into better diagnosis, treatment and prevention of cancer.

INTRODUCTION

In this past year, remarkable advances in our understanding of the mutational profiles and other disease-specific alterations of cancer genomes have been reported. In general, the field of cancer genomics has been impacted most profoundly by the application of next-generation sequencing technology, which has tremendously accelerated the pace of discovery while dramatically reducing the cost of data production. Hence, there has been a rapid progression from targeted gene re-sequencing using PCR and Sanger sequencing to either targeted, whole genome, or whole transcriptome sequencing using these massively parallel sequencing platforms, coupled with the requisite bioinformatics-based approaches to analyze the data. Within this brief timeframe, studies examining all known genes in a few samples to those examining hundreds of genes in hundreds of samples, to whole genome sequencing and analysis of a matched tumor/normal pair have been reported. There remains much to be learned about this complex disease, of course, but our fundamental understanding of which genes are mutated in cancer cells, the pathways that are impacted by these mutations, and how these data inform our models of cancer biology will undoubtedly evolve rapidly in the near future.

STRUCTURAL VARIATION STUDIES

A well-known characteristic of cancer genomes is that they are frequently altered in their gross chromosomal structure by amplification, deletion, translocation and/or inversion of chromosomal segments. Such alterations often, of course, concomitantly alter genes in a number of ways that may be critical to cancer onset or progression. As such, important developments in obtaining increasingly more detailed genome-wide characterizations of structural variation (SV) in tumor genomes have been described recently. Initially, these studies were conducted using signal strength-based analyses on high-density SNP array data sets, where tumor and normal genomic DNA were compared and any large-scale amplification or deletion signals were detected as continuous blocks of SNPs with higher than (amplification) or lower than (deletion) the normalized signal strength (1). The genes in these regions often are re-sequenced to identify mutations or are assayed for evidence of altered gene expression levels that correlate with a detected copy number alteration. Weir *et al.* (2) provided a powerful example of this approach using 384 lung adenocarcinoma samples in which they identified a novel candidate proto-oncogene (NKX2-1/TITF1) in an amplified region of chromosome 14.

Complementary to array-based methods, next-generation sequencing-based approaches are being applied to the SV problem at a higher level of resolution and complexity. Korbel *et al*. (3) first demonstrated that paired-end reads from next-generation sequencing platforms can be aligned to the genome and examined algorithmically to identify putative SV. Their approach was based on the identification of anomalously mapping read pairs that align several standard deviations outside the well-defined size range of the library itself. Read pairs that mapped too close together, too far apart, in an unpredicted orientation, or across chromosomes gave the indication of potential insertions, deletion, inversions or translocations in the sequenced genome. By these methods, we can obtain a much more precise view of genome-wide SV than by array-based analysis methods. Several groups have

*To whom correspondence should be addressed at: Department of Genetics, Washington University School of Medicine, 4444 Forest Park Boulevard, St. Louis, MO 63108, USA. Tel: +1 3142861805; Fax: +1 3142861810; Email: emardis@wustl.edu recently described advanced implementations of this approach; utilizing low coverage of a cancer genome with paired end reads (4,5). These methods fit nicely into a paradigm of whole genome sequencing followed by mutation discovery. Here, a small investment in paired end reads at light coverage can profile the extent of SV across a large number of tumor samples as a first step. This type of analysis not only identifies common copy number and structural variant loci, but can also allows a calculation of the deeper sequence coverage that will be required to characterize focal mutations (e.g. single nucleotide and small in/dels) in each tumor genome, since large-scale amplification (for example) will inflate the sequence coverage requirement. One can then obtain this deeper coverage with the same libraries used to produce the initial data set.

TARGETED GENE SEQUENCING STUDIES

The combination of PCR and Sanger sequencing to discover mutations in tumor genomes has proven a powerful initial approach, as evidenced by several recent studies that we describe below. Although studies using this method have targeted limited numbers of genes and successfully identified key somatic mutations in cancer genomes, the method recently has been applied to characterize hundreds of genes as well as the entire ‘exome’ (all known protein coding exons). In particular, two articles published in the same 2008 issue of Nature (6,7) demonstrated how targeted gene re-sequencing and variant detection can contribute significantly to our understanding of the types of genes carrying somatic mutations in a given cancer type [here, lung adenocarcinoma and glioblastoma multiforme (GBM)] by discovering novel genes mutated in each tumor type. In these studies, by virtue of sequencing large numbers of the same tumor type (based on pathological examination, tumor stage and grade), the results highlighted the cellular pathways putatively impacted by these mutations. Both articles arrived at important correlative conclusions by integrating the somatic mutation data with the results from other genome-wide characterizations of the same samples, such as array-based gene expression data, genome structure perturbation data [e.g. loss-of-heterozygosity (LOH), amplification or deletion of large chromosomal segments], and clinical data elements (e.g. outcome, response to therapy, etc.). For example, MAPK signaling, P53 signaling, cell cycle regulation and mTOR pathways are targeted in lung adenocarcinoma samples by combinations of point mutation, copy number amplification and deletion and LOH (7).

Similarly, Vogelstein and colleagues have extended their initial efforts to characterize mutations by screening most of the known coding genes in the genome in several tumor types (8,9), to also include information about gene expression using next-generation sequencing of serial analysis of gene expression tags, and about genome copy number alterations from genotyping arrays. Their analyses combine data about somatically mutated genes with data about copy number alterations to identify candidate cancer genes (‘CAN-genes’), thereby generating evidence for mutations that are driving carcinogenesis (‘drivers’) versus having no impact on tumor growth (‘passengers’). Gene expression data inform the pathways analysis, by reflecting epigenetic alterations not detectable by sequencing or copy number analyses.

This combined approach, in a study of GBM samples, resulted in the discovery of several commonly mutated genes, some impacting novel pathways. Among these was the surprising identification of an IDH1 mutation that was found in 18/149 (12%) cases, all occurring at the same residue (R132) (10). Using clinical data, several interesting correlations regarding the IDH1 mutation were made; namely that this mutation was more prevalent in younger GBM patients (mean age of 33 versus 53 years of age), more prevalent in patients developing secondary GBMs (that develop from low grade gliomas) and predicted a significantly improved prognosis (median overall survival of 3.8 versus 1.1 years). In a follow-on study, this group evaluated the IDH1 R132 and related IDH2 R172 mutation prevalence in a much wider range of tumor types that included 445 central nervous system (CNS) tumors and 494 non-CNS tumors (11). Here, the previously observed improved outcome for GBM patients carrying the IDH1 mutation was confirmed and extended to those carrying mutated IDH2 (median overall survival of 31 versus 15 months, at $P = 0.002$), and for patients with anaplastic astrocytomas (median overall survival of 65 versus 20 months, $P < 0.001$). An evaluation of the impact of one IDH1 mutation (R132H) and three IDH2 mutations (R172G, K and M) on the function of the resulting proteins showed severely diminished activity in NADPH production relative to the wild-type enzymes.

TRANSCRIPTOME CHARACTERIZATION

As more detailed profiling of the cancer genome has developed, the need for a full understanding of how these somatic alterations are manifest in the genes expressed by tumors has become pertinent. As in genome characterization, the use of next-generation sequencing of RNA extracted from tumor cells (‘RNA-seq’) produces a comprehensive data set for complete transcriptome characterization, as well as correlation to known genomic changes such as structural and copy number alterations, focused in/dels and single nucleotide mutations. Not only does this approach greatly expand the dynamic range of gene expression level data beyond the sensitivity limits of microarrays (12), but also it provides data that can be further mined in a number of ways (13) to enhance the understanding of the transcriptome in cancer. For example, RNA-seq data can identify allele-specific expression in the context of known mutations, verify the impact of a nonsense mutation, or provide a means of finding mutations in tumors as illustrated recently in ovarian tumors (14). Here, four granulosa-cell tumors (GCT) of the ovary were analyzed using whole transcriptome paired-end RNA sequencing, demonstrating that all four GCTs had a missense point mutation in the FOXL2 gene. This gene encodes a transcription factor known to be crucial in granulosa cell development, and since the same mutation was determined to be present in additional GCTs of the same adult-type tumors, it is a potential driver mutation.

These data also can be analyzed to detect alternative splice isoforms and fusion transcripts (15), as illustrated recently in a very clever approach by Maher *et al*. (16) that identified both known and novel fusion transcripts in prostate cancer samples. This approach utilized a combination of two next-generation platforms to produce sequence reads that were combined to identify fusion transcripts from cancer cell lines. In particular, RNA-seq data from a longer-read technology (Roche/454) first identified putative fusion transcripts by virtue of their alignment characteristics to the transcriptome, and then a second RNA-seq data set from short read length platform (Illumina Genome Analyzer) was aligned to the putative fusion transcript reads to provide support for their presence. Using this paradigm, Maher *et al*. successfully identified known and novel fusion transcripts in the prostate cancer cell lines LnCaP and VCaP, and subsequently in RNA from several prostate tumor samples.

RNA-seq also can build evidence for novel genes that previously have not been annotated due to lack of ESTs or were missed by *in silico* prediction (13,17). Hence, further development of methods that elucidate the complexity of the transcriptome in cancer will both support and enrich our understanding of the cancer genome and cancer biology.

In addition to mRNA, the study of microRNAs (miRNAs) and their roles in regulating the expression of specific genes in both healthy and cancerous cells is rapidly expanding our comprehension about this aspect of cell biology (18). A recent study by Uziel *et al*. (19) demonstrated the interaction between miRNA overexpression and a well-characterized signaling pathway, Sonic Hedgehog/Patched (SHH/PTCH) in medulloblastoma (MB). Having determined the overexpression of nine genes in the miR-17–92 cluster in an MB mouse model with constitutively activated SHH/PTCH signaling pathway, this group then tested and demonstrated similar miR-17–92 cluster upregulation in a subset of human MB tumors with constitutively activated SHH/PTCH. This study provided the first evidence that the SHH/PTCH signaling pathway and miR-17–92 functionally interact and contribute to both murine and human MB development.

Similarly, Wyman *et al*. (20) and Nygaard *et al*. (21) demonstrated detection of novel miRNAs and miRNAs with differential expression in ovarian and breast cancer, respectively, using Roche/454 sequencing and miRNA discovery bioinformatics pipelines. Building upon these studies and others, numerous groups are now proposing miRNAs as prognostic or diagnostic markers for a variety of cancer types (22–25).

WHOLE GENOME SEQUENCING

The most significant impact of next-generation sequencing on cancer genomics has been the ability to re-sequence, analyze and compare the matched tumor and normal genomes of a single patient. With the significantly reduced cost of sequencing and tremendously enhanced throughput, it is now within the realm of possibility to sequence multiple patient samples of a given cancer type. Such efforts require not only data generation, but also the careful development of analytical tools and pipelines, supported by validation efforts that feedback into the analytical process, to enhance the sensitivity and specificity of variant discovery. Due to the complex nature of genome variation, the entire spectrum of potential mutations requires consideration, including germline susceptibility loci, somatic single nucleotide and small indel mutations, copy number alterations and structural variants. To-date, one publication has outlined such a study, describing the results obtained from sequencing and analysis of an acute myeloid leukemia genome (26). Several key concepts have emerged from this approach, including the use of high-density SNP genotype data to estimate genome sequence coverage by tracking the accuracy of sequence-based SNP calls at heterozygous loci, a step-wise approach to somatic single nucleotide variant discovery, and the use of read counts to establish the prevalence of somatic variants in the tumor cell population. The basic analytical approach aligned tumor (~21-fold haploid coverage) and normal (~14-fold haploid coverage) sequence reads to the reference human genome using the *Maq* alignment algorithm (27). As coverage accumulated during the generation of tumor and germline reads, *Maq* was used to call variant positions across the genome, and those calls were compared with the heterozygous loci determined from the overlapping set of SNP array genotype calls identified by both Illumina and Affymetrix genotyping arrays. Sequence coverage was considered sufficient for mutation discovery once heterozygous calls from sequence data were made for >95% of these orthogonally determined heterozygous SNP positions. This approach toward monitoring genome coverage is now a cornerstone of our cancer genome re-sequencing pipeline.

Somatic mutation discovery requires a number of steps to eliminate from consideration all known sequence variants, typically by (1) comparison with other sequenced genomes (via dbSNP) and to other resources for variant discovery such as the 1000 Genomes Project (www.1000genomes.org), followed by (2) comparison at remaining variant sites between the tumor and the normal genome. The approach also takes into consideration two primary measures of quality in order to distinguish high- from low-quality variants in the latter comparison. These primary measures include first, a cumulative base-calling quality value that is summed from the individual quality values of each base identifying the putative variant (assigned by the Illumina analysis pipeline) and second, a mapping quality value assigned by Maq that indicates the genome-wide uniqueness of each aligned read. Nonetheless, false positives do occur in this analysis, as do false negatives. False positives tend to result from incorrect interpretation of one or more data elements considered by the multicomponent analysis algorithm, often due to non-unique read placement or to a missing variant call in the matched normal sequence. The false negatives are harder to evaluate, but mainly appear to be due to lack of sufficient read support for a true variant in the tumor. On one hand a reasonably high false positive rate is desired so true mutations are not missed, but on the other it is important to known which predictions are incorrect. Because of this, performing an orthogonal validation step using PCR-directed sequencing or genotyping to establish false from true positives for all putative somatic variants in genes or in regulatory/conserved regions of the genome should be done.

One of the key aspects of evaluating somatic mutations in cancer genomes is that the collective sequencing read pool

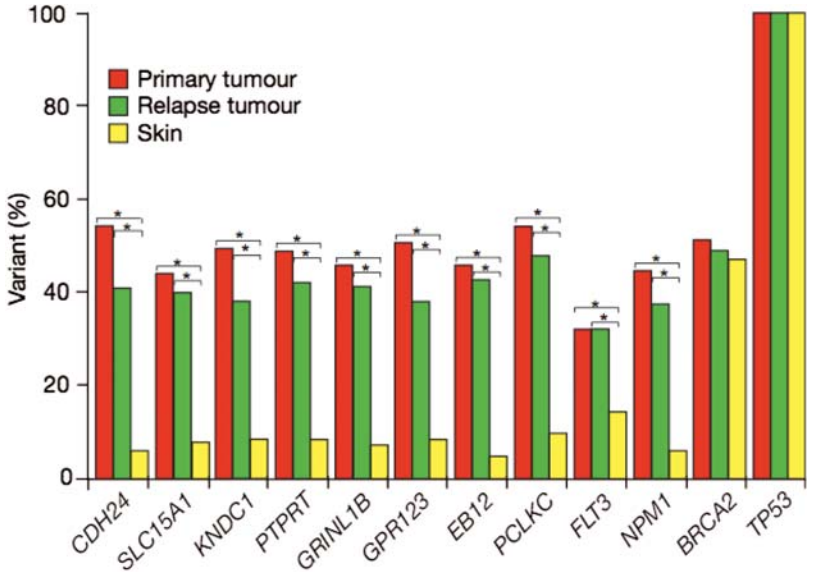

Figure 1. Summary of readcount data obtained for ten somatic mutations and two validated SNPs in the AML primary tumor, AML relapse tumor and normal skin specimens. As described in the text, all heterozygous mutations were determined by readcount data to be present at around 50% prevalence in the tumor cells with the exception of the FLT3 internal tandem duplication mutant. The variant alleles in the primary and relapse tumor samples are statistically different from that of the skin sample for all mutations. Note that the normal skin sample was contaminated with leukemic cells containing the somatic mutations, as the patient's white blood cell count was 105 000 when the skin punch biopsy was obtained.

represents a census of the genomic DNA contributed from all cancer cells used for DNA isolation. One challenge of this pooled approach is to determine what proportion of those cells carried each identified mutation. Information about the prevalence of any mutation in a cell population allows one to infer how early in the path toward cancer development that particular mutation occurred. The digital nature of next-generation sequencing allows us to evaluate this prevalence, since each read in the sequenced pool of fragments represents a single original DNA fragment from that cancer cell census. For example, since many mutations will present as heterozygous, we expect that 50% of the reads in a pure tumor cell population will contain the variant. Obviously, this proportionality will be influenced by the percentage of tumor cells in a sample, so a correction factor is applied based either on estimates from pathology review or by a more precise measure that calculates the percentage of normal reads present in the tumor read population at known/validated somatic sites in that tumor genome (L. Ding, personal communication). This type of analysis was applied to the first AML genome sequence, demonstrating that all somatic mutations were found in virtually all of the cells of the tumor, except for the FLT3 internal tandem duplication (Fig. 1), which is known from mouse models to not be an initiating mutation in AML (28).

We recently published our findings from sequencing a second AML genome and matched normal (29), where we employed the aforementioned concepts, identifying nine single nucleotide somatic variants in genes, two genic indels, and 54 somatic single nucleotide variants in known regulatory or highly conserved regions of the genome. Although none of the novel somatic variants identified in the first AML genome were recurrent among 187 other AML tumor genomes tested, one mutation found in the second AML genome analysis proved to be recurrent in 8.2% of those samples. This gene was IDH1, mutated at the exact R132 site also identified in GBM (10), as described earlier. Unlike Parsons *et al.*, however, our correlation analysis among the 187 AML patients, combined with the clinical data, indicated that in AML, the IDH1 mutations portend a significantly worse outcome by Kaplan–Meier analysis for those patients who have normal cytogenetics and lack the NPMc and FLT3 mutations (Fig. 2). This finding demonstrates the power of the genomics approach, and highlights how new insights into cancer biology will result from further cancer genome sequencing.

CANCER GENOME SEQUENCING: THE FUTURE

One clear trend in cancer genome sequencing is that the continuing advance of next-generation technology in terms of data capacity per instrument run and read length will accelerate the rate of sequencing whole genomes, at ever-decreasing costs. Since next-generation platforms can produce data to characterize gene expression, methylation, histone packaging, transcription factor and other regulatory protein binding positions, and so on, we can build data sets that quite comprehensively characterize a broad spectrum of genomic alterations among sets of tumor samples.

A key question is what the planned sequencing of hundreds of tumors might reveal? For example, it is not yet clear whether the cancer-critical somatic alterations we identify will be found to recurrently affect specific genes, or if the combination of recurrent and 'private' mutations will define each cancer genome and hence, its treatment. We also need to understand the potential role of inherited genomic variation in shaping the onset of cancer and its outcomes, which is one reason sequencing a matched normal sample from each patient

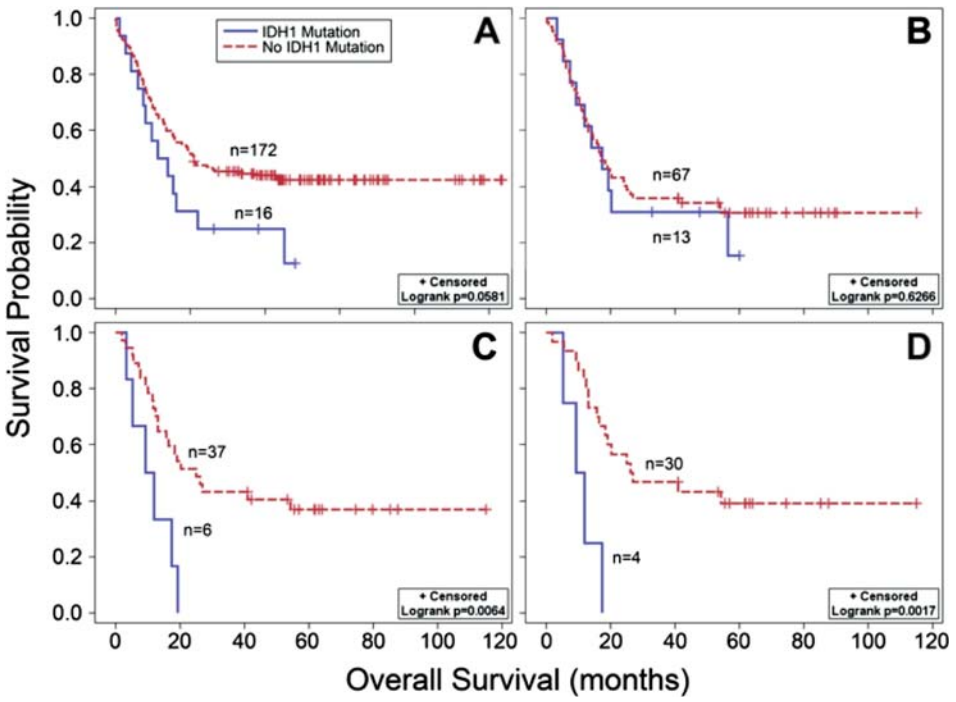

Figure 2. Kaplan–Meier survival analysis of AML patients. Overall survival is shown for patients with (blue) or without (IDH1) mutations. (**A**) Entire AML cohort. (**B**) AML patients with normal cytogenetics. (**C**) AML patients with normal cytogenetics and non-mutated NPM1. (**D**) AML patients with normal cytogenetics, non-mutated NPM1 and non-mutated FLT3. Differences were assessed by Log rank analysis.

is so important. Determining the genomic landscape of hundreds of tumors ultimately will dictate whether each cancer genome will require a full genome variation profile as a diagnostic component of individualized treatment. It is imperative also to focus some genome characterization efforts toward elucidating the genomic changes that distinguish primary from metastatic disease.

Once we understand the genomic landscape of cancer, what should follow? Whereas genome-wide characterization of tumors likely will yield important clues about the genes that play a role in carcinogenesis or metastasis, we must be prepared to follow-up on these clues by carrying out functional screens of altered genes with commensurately high-throughput capabilities. Functional screening would aim to identify those somatic alterations that are initiating carcinogenesis, or promoting metastasis, thereby establishing candidate genes and their protein products for targeted therapy development or testing, as well as for diagnostic/prognostic assay development. Luo *et al*. (30) have published such one approach, employing pooled short hairpin RNA (shRNA) screening paradigms of cancer cell lines that identified genes essential for growth and related phenotypes in these cells, as well as genes involved in the response of cancer cells to tumoricidal agents. Lynda Chin and colleagues (31) recently published an elegant example of a complete genomics-to-function paradigm, first identifying a genomic region at 5p13 that was commonly amplified in several cancer types (lung, ovarian, prostate, breast, melanoma), and then using integrated analysis of this region to pinpoint the Golgi-associated protein GOLPH3 for further study. Using a variety of clues from the results of *in vitro* shRNA knock-down of GOLPH3 in cell lines that either did or did not contain the 5p13 amplification, to *in vivo* GOLPH3 overexpression in these same cell lines, to clues from yeast genetics that linked GOLPH3 to the *trans*-Golgi network and ultimately as a determinant of rapamycin sensitivity as a regulator of mTOR, the study established GOLPH3 as a first-in-class Golgi oncoprotein. This result further emphasizes the need for multiple lines of evidence to support functional and mechanistic roles for the genomic alterations we are finding in cancer genomics today.

ACKNOWLEDGEMENTS

The authors wish to acknowledge Drs Devin Locke and Li Ding, for their critical reading of the manuscript.

Conflicts of Interest statement. None declared.

FUNDING

We thank the National Human Genome Research Institute for support of this research via U54-HG003079 (R.K.W.).

REFERENCES

1. Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J.C., Huang, J.H., Alexander, S. *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA*, **104**, 20007–20012.
2. Weir, B.A., Woo, M.S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W.M., Province, M.A., Kraja, A., Johnson, L.A. *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature*, **450**, 893–898.

3. Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L. *et al.* (2007) Paired-end mapping reveals extensive structural variation in the human genome. *Science*, **318**, 420–426.
4. Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P. *et al.* (2009) BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat. Methods*. Epub ahead of Print.
5. Chiang, D.Y., Getz, G., Jaffe, D.B., O'Kelly, M.J., Zhao, X., Carter, S.L., Russ, C., Nusbaum, C., Meyerson, M. and Lander, E.S. (2009) High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat. Methods*, **6**, 99–103.
6. Consortium, T.C.G.A. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
7. Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B. *et al.* (2008) Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, **455**, 1069–1075.
8. Wood, L.D., Parsons, D.W., Jones, S., Lin, J., Sjoblom, T., Leary, R.J., Shen, D., Boca, S.M., Barber, T., Ptak, J. *et al.* (2007) The genomic landscapes of human breast and colorectal cancers. *Science*, **318**, 1108–1113.
9. Sjoblom, T., Jones, S., Wood, L.D., Parsons, D.W., Lin, J., Barber, T.D., Mandelker, D., Leary, R.J., Ptak, J., Silliman, N. *et al.* (2006) The consensus coding sequences of human breast and colorectal cancers. *Science*, **314**, 268–274.
10. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C., Leary, R.J., Angenendt, P., Mankoo, P., Carter, H., Siu, I.M., Gallia, G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
11. Yan, H., Parsons, D.W., Jin, G., McLendon, R., Rasheed, B.A., Yuan, W., Kos, I., Batinic-Haberle, I., Jones, S., Riggins, G.J. *et al.* (2009) IDH1 and IDH2 mutations in gliomas. *N. Engl. J. Med.*, **360**, 765–773.
12. Marioni, J.C., Mason, C.E., Mane, S.M., Stephens, M. and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.
13. Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.
14. Shah, S.P., Kobel, M., Senz, J., Morin, R.D., Clarke, B.A., Wiegand, K.C., Leung, G., Zayed, A., Mehl, E., Kalloger, S.E. *et al.* (2009) Mutation of FOXL2 in Granulosa-Cell Tumors of the Ovary. *N. Engl. J. Med.*, **360**, 2781–2783.
15. Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.
16. Maher, C.A., Kumar-Sinha, C., Cao, X., Kalyana-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. and Chinnaiyan, A.M. (2009) Transcriptome sequencing to detect gene fusions in cancer. *Nature*, **458**, 97–101.
17. Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.
18. He, L., He, X., Lowe, S.W. and Hannon, G.J. (2007) microRNAs join the p53 network—another piece in the tumour-suppression puzzle. *Nat. Rev. Cancer*, **7**, 819–822.
19. Uziel, T., Karginov, F.V., Xie, S., Parker, J.S., Wang, Y.D., Gajjar, A., He, L., Ellison, D., Gilbertson, R.J., Hannon, G. *et al.* (2009) The miR-17–92 cluster collaborates with the Sonic Hedgehog pathway in medulloblastoma. *Proc. Natl. Acad. Sci. USA*, **106**, 2812–2817.
20. Wyman, S.K., Parkin, R.K., Mitchell, P.S., Fritz, B.R., O'Briant, K., Godwin, A.K., Urban, N., Drescher, C.W., Knudsen, B.S. and Tewari, M. (2009) Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. *PLoS ONE*, **4**, e5311.
21. Nygaard, S., Jacobsen, A., Lindow, M., Eriksen, J., Balslev, E., Flyger, H., Tolstrup, N., Moller, S., Krogh, A. and Litman, T. (2009) Identification and analysis of miRNAs in human breast cancer and teratoma samples using deep sequencing. *BMC Med. Genomics*, **2**, 35.
22. Hu, X., Macdonald, D.M., Huettner, P.C., Feng, Z., El Naqa, I.M., Schwarz, J.K., Mutch, D.G., Grigsby, P.W., Powell, S.N. and Wang, X. (2009) A miR-200 microRNA cluster as prognostic marker in advanced ovarian cancer. *Gynecol. Oncol.*, **114**, 457–464.
23. Aqeilan, R.I., Calin, G.A. and Croce, C.M. (2009) miR-15a and miR-16-1 in cancer: discovery, function and future perspectives. *Cell Death Differ.*, Epub.
24. Busacca, S., Germano, S., De Cecco, L., Rinaldi, M., Comoglio, F., Favero, F., Murer, B., Mutti, L., Pierotti, M. and Gaudino, G. (2009) MicroRNA Signature of Malignant Mesothelioma with Potential Diagnostic and Prognostic Implications. *Am. J. Respir. Cell Mol. Biol.*, Epub.
25. Dyrskjot, L., Ostenfeld, M.S., Bramsen, J.B., Silahtaroglu, A.N., Lamy, P., Ramanathan, R., Fristrup, N., Jensen, J.L., Andersen, C.L., Zieger, K. *et al.* (2009) Genomic profiling of microRNAs in bladder cancer: miR-129 is associated with poor outcome and promotes cell death in vitro. *Cancer Res.*, **69**, 4851–4860.
26. Ley, T.J., Mardis, E.R., Ding, L., Fulton, B., McLellan, M.D., Chen, K., Dooling, D., Dunford-Shore, B.H., McGrath, S., Hickenbotham, M. *et al.* (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. *Nature*, **456**, 66–72.
27. Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.
28. Kelly, L.M., Kutok, J.L., Williams, I.R., Boulton, C.L., Amaral, S.M., Curley, D.P., Ley, T.J. and Gilliland, D.G. (2002) PML/RARalpha and FLT3-ITD induce an APL-like disease in a mouse model. *Proc. Natl. Acad. Sci. USA*, **99**, 8283–8288.
29. Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D. *et al.* (2009) Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *N. Engl. J. Med.*, Epub.
30. Luo, B., Cheung, H.W., Subramanian, A., Sharifnia, T., Okamoto, M., Yang, X., Hinkle, G., Boehm, J.S., Beroukhim, R., Weir, B.A. *et al.* (2008) Highly parallel identification of essential genes in cancer cells. *Proc. Natl. Acad. Sci. USA*, **105**, 20380–20385.
31. Scott, K.L., Kabbarah, O., Liang, M.C., Ivanova, E., Anagnostou, V., Wu, J., Dhakal, S., Wu, M., Chen, S., Feinberg, T. *et al.* (2009) GOLPH3 modulates mTOR signalling and rapamycin sensitivity in cancer. *Nature*, **459**, 1085–1090.